基于医案元数据的中医诊疗数据集构建方法与实证研究

马捷[1,2] 王珏[1] 孙恒宇[3] 李璐[1]

[1] 吉林大学管理学院 长春 130022 [2] 吉林大学信息资源研究中心 长春 130022

[3] 吉林省中医药科学院第一临床医院 长春 130022

摘 要:[目的/意义]中医医案是记录诊疗数据的重要载体,基于中医医案元数据进行中医诊疗数据采集,形成数据集,对于中医诊疗知识共享、挖掘和传承具有重要意义。[方法/过程]选择和确定中医医案元数据,并参考关系数据库对数据集进行概念与逻辑设计,提出对中医医案数据进行采集、组织和存储,进而形成数据集的方法,并进行实证研究。[结果/结论]基于较完善的中医医案元数据进行网络和实地诊疗数据采集,形成的中医诊疗数据集不仅可以支持多维度的诊疗信息存储、分享与检索,也能为更深入的诊疗数据挖掘提供数据源。

关键词:医案 中医医案元数据 中医诊疗数据集

分类号:G251.2

DOI:10.13266/j.issn.0252-3116.2021.02.003

中医经过千年的发展和沉淀,形成了一套自己独有的思维模式,但其传播和发展仍然以“言传身教”为主,且由于中医自身“以人为本”“千人千方”等特点,并未形成一套广泛使用的诊疗模式,使得中医诊疗阶段大量高价值数据无法很好地保存与分析,限制了中医的发展。究其原因,主要是因为信息传播不畅和交流障碍,而特定主题数据集的建立与高质量科学数据的出版是达到数据共享、打破信息孤岛的重要手段[1]。因此,提出一套科学记录中医诊疗细节的元数据,并基于此元数据建立中医诊疗数据集,将中医诊疗过程中的大量数据记录并保存下来,进行数据分析与隐性知识挖掘,可以极大推动中医诊疗信息的开发和利用。

中医诊疗所涉及知识纷繁复杂,为了保证数据集质量,应选择合理的采集领域。章太炎先生说:“中医之成绩,医案最著”[2],中医医案是中医理、法、方、药综合运用的具体反映形式,它不仅真实记述了医疗活动,同时也反映了医家的临床经验及诊疗思维[3],最重要的是详细记录了患者的信息,没有只考虑疾病本身。因此,历代医案、中医网络医案以及临床记录的医案,都是数据集进行数据采集的数据源。然而无论上述哪一种医案,其记录一般都是以叙述性的非结构化文本为主,医案的字段划分粒度较大,无法展现中医医案细节,不利于中医诊疗知识挖掘。因此,对非结构化文本进行结构化处理就显得非常必要,一是进行粒度细化,合理拟定元数据,将医案文本中的重要信息结构化;二是进行术语标准化处理,即将文本中医疗术语(如药品、疾病名称)映射至权威或标准化的形式[4],这样不仅可以减少诊疗过程记录的缺陷率与错误率,提高数据质量,更有利于专业知识的共享与交流[5]。

综上,本文提出基于中医医案元数据,构建高质量中医诊疗数据集,以期实现中医诊疗数据的结构化存储,支持深度检索和数据挖掘。

1 文献综述

从上世纪90年代以来,不同应用领域中有不同的元数据标准出现。为了规范信息存储,方便资源的共享和利用,国内外相关领域内对科学数据的组织多采用元数据方式,但元数据模型较多,即使在同一领域内也没有公认的元数据模型可以满足所有应用[6]。根据贾李蓉等[7]对国际标准化组织(ISO)发布的元数据标准研究,目前尚未有一套体系完整、使用广泛的中医领域元数据,需要根据具体使用情况对已有元数据进行扩展。赵阳等[8]通过对中医药文献元数据的著录对象的界定和中医文献元数据的必要性分析,规划元数据

作者简介:马捷(ORCID:0000-0002-1471-2143),教授,博士,博士生导师;王珏(ORCID:0000-0001-5272-7904),硕士研究生,通讯作者,E-mail:1069821287@qq.com;孙恒宇(ORCID:0000-0003-0882-7507),主治医师;李璐(ORCID:0000-0003-0460-1789),硕士研究生。

收稿日期:2020-05-19 修回日期:2020-10-02 本文起止页码:27-36 本文责任编辑:易飞

框架结构并提出元数据来源和扩展原则，为中医领域元数据拟定提供了思路。刘宝杰等[9]在参阅了大量病案的基础上，设计了一套中医病案数据库的元数据方案，元数据方案分为医案基本信息、疾病基本信息、诊断、治疗、方药、错误操作、禁忌、西医相关信息、医嘱 9 个部分，初步达到了组织和描述中医病案资源的目的，但却存在“全而不精”的问题，过于笼统的字段表述以及相关标准的缺失会造成收集数据的错位与混乱。孙静等[10]以本体论为基础，探讨了中医医案信息的信息获取与管理方法，利用本体来组织数据，可以很清晰地揭示数据之间的内在联系，很大程度上弥补元数据对于知识组织存在的不足，有利于构建知识共享系统，但该研究对于中医医案的挖掘相对较浅，字段划分粒度较大，对于临床医案应用有限。

数据集是具有一定主题、可以被标识并能够被计算机处理的数据集合，是一种基于主题进行数据资源收集与组织的新型数据组织方式和数据资源的“封装”单元[11]。构建专业领域内的数据集是目前被广泛采用的对数据资源进行收集、整合与检索的方法。目前构建数据集的主要方式为对领域内文献进行数据的遴选[12]或对相关主题数据进行结构化采集，如单连慧等[13]对 PubMed 和 SCIE 两个平台数据库的相关数据进行获取和整合，最终基于 DOI 构建了医学领域科技评价文献数据集，但该方法也受数据库以及检索主题词选择的影响，其准确性与全面性只能得到初步验证，且面对医学其他领域的应用也待进一步研究；邱瑞瑾等[14]通过对相关领域进行文献研究和采用半结构化访谈的方法形成临床安全性评价核心数据原始项目清单，并利用德尔菲调查法对纳入核心数据集的数据进行修正，最终通过共识会议形成了核心数据集，这种方法尽可能多地考虑了相关因素的影响，但对于核心数据的选择可能无法达成完全共识。对于医学数据来说，其关联数据范围十分广泛，涉及到疾病治疗、公共卫生、人口统计等多个方面[15]，因此构建医学数据集要对应用场景以及主题外延进行合理限定，避免一味求全造成主题不够明确，导致数据爆炸；此外还要综合考虑构建者及领域专家的主观局限性等因素影响，多运用访谈以及德尔菲调查法进行数据的遴选与修正。目前国内外医学领域数据集的建立对健康管理、医学影像管理、疾病预测等具有重要的数据支持和辅助决策作用。全球已经建立起了 OASIS、NSCLC 等多个医学病例分析数据集，此外还有医学降噪数据集与医学分割数据集等，其中大部分为图像数据集，如大脑结构数据集、乳腺数据集等。在记录临床信息方面，比较著名的有 MIMIC 数据集[16]，它用来记录病人的临床信息，包含人口统计学、床边生命体征测量、护理人员笔记、出院信息以及相关的影像学报告。MIMIC 支持涵盖流行病学、临床决策规则改进和电子工具开发的各种分析研究。此数据集体量庞大，几乎包含一位患者可以涉及的所有静态数据，虽然其对诊疗构成的记录方式更适用于西医，但其整体结构对于此次中医数据集构建依然有很大借鉴意义。

为了推进中医领域资源共享，我国已建成多个用于中医数据管理与研究的知识库。如中国知网中医药知识资源总库、疾病诊疗知识库、万方医学网临床诊疗知识库等，其中收录了丰富的疾病、检查、药品、循证、病例、文献等资源，为临床决策提供了丰富可靠的支持。一个标准的知识库应具有推理机制，能实现知识推理与知识挖掘，但部分知识库由于其事实数据层存储混乱，没有合理的数据表结构，且元数据字段划分不够科学，不能体现中医诊疗特色，造成数据冗余、数据冲突，其功能仍停留在数据库层面，本质上还是不同领域的数据集合。无论构建中医医案数据集还是知识库，所应用的元数据要能够使具有诊疗价值的信息充分结构化，还要做好术语的规范化和标准化工作。

2 中医医案元数据选择与术语规范化处理

2.1 元数据的选择

元数据是描述数据属性的结构化数据，功能主要为资源的组织、挖掘、互操作、数字鉴别和保存[17]。如果临床数据大多处于散乱、无序的状态，并且存在记录格式不统一、存档方式无统一规范，那么就无法有效地共享与查找资源，造成大量临床数据流失，形成信息孤岛[18]。目前，国际上还没公认的中医病案元数据标准。因为中医语言特点，国际通用的元数据标准虽然有很好的交互性，但并不能准确描述中医资源，而中医医案记录的行文自由性也局限了元数据标准的选择。本研究团队李璐等整理设计了一套中医医案元数据[19]，并在综合分析大量标准以及文献的基础上，运用内容分析法、专家咨询法和实地调研法进行完善，最终得到一套比较合理、普遍适用且精度较高的中医医案元数据方案。本研究数据集构建即以此套元数据为蓝本，在原有的元数据结构上做了一些细微调整，以便更好地适应数据的采集与保存。

2.2 元数据结构的修订

为了让医学工作者方便快捷地输入数据，让数据

管理者可以对数据集内的数据进行筛查整理，让导出的数据质量满足科研工作者的分析要求，根据元数据的功能和特点，拟将整体元数据分为两大结构：医案描述元数据与医案管理元数据。其中，医案描述元数据是详细记载诊疗过程的元数据；医案管理元数据则对医案本身的保存流转进行客观记录。通过对网络上现有医案病历进行分析可以得出以下几种情况：①出于对个人隐私的保护，患者的基本信息部分常常缺省；②患者主诉症状往往与四诊情况一起描述，即由患者口述主要病情后，医生随即通过望、闻、问、切四诊合参来进一步辩证分析；③医生诊断与药方医嘱间联系紧密。因此，平衡各部分的信息量，考虑医案整体脉络以及内在逻辑，将整个医案描述元数据部分分为医案标识信息集、患者基本信息集等5个子集；再从医案保存管理的角度考虑，参照现阶段医院门诊病历的存放规则，建立医案管理、流转信息子集，记录医案保存传递的具体信息。具体结构如图1所示：

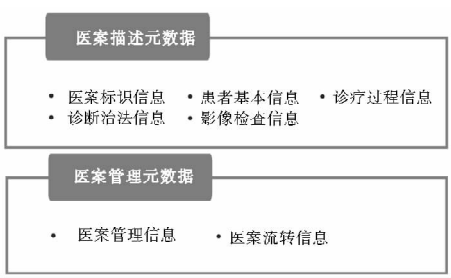

图1 元数据结构

厘清元数据结构之后，对每种元数据的每张信息表进行简单的字段划分，为后续设计做好准备，如表1所示：

表1 中医诊疗数据集各信息表元数据情况汇总

信息表	字段	元数据数量/个	元数据总量/个
医案标识信息表	–	8	8
患者基本信息子集	基本信息字段	8	24
	体格检查字段	4	
	辅助检查字段	1	
	既往病史字段	8	
	个人及家族史字段	3	
诊疗过程信息子集	主诉字段	3	24
	四诊合参字段	21	
诊断治法信息子集	诊断字段	4	12
	治法字段	8	
影像检查信息子集	–	6	6
医案流转信息子集	–	12	12
医案管理信息子集	–	5	5
总计			91

2.3 术语规范化处理

在以上各个信息子集中，由于所涉及的数据内容、数据来源以及用途各不相同，所以各个子集内所选用的标准也不尽相同，而标准化术语是医学信息学发展的基础，它为记录信息提供了基本的框架结构，使原始信息数据保持一致，是解决语义互操作性、知识表达一致性以及医疗资源共享性的有效手段[12]。在中医领域，《中医药学语言系统语义网络框架》规定了中医药学语言系统的语义类型、语义概念及语义关系，并对其进行了详细定义。该标准的提出不仅规范和支持了中医药学语言系统的建设，还为中医药学术语系统和本体创建提供了语义标准，为中医药学语言系统和统一的医学语言系统映射提供了支持，对中医药学术语信息的交换具有重要意义[20]。根据医学领域对术语标准化的规定，医学名词应使用全国科学技术名词审定委员会公布的名词，另外对尚未确定使用标准的名词参照表2执行：

表2 医学领域不确定名词使用标准

主题词表及术语系统	适用情况	备注
医学主题词表（MESH）	尚未通过审定的学科名词	标准医学主题词表
医学主题词注释字顺表	尚未通过审定的学科名词	是MESH的扩充版，专供标引、编目和联机检索使用
中国中医药学主题词表	尚未通过审定的学科名词	为建立中医药科技情报检索体系奠定基础，以适应中医药事业的发展
中医临床诊疗术语疾病、证候、治法部分	中医名词术语	用于中医病案的用词规范，病案中中医疾病诊断信息的统计
中华人民共和国药典	尚未有通用译名的名词术语	简称《中国药典》，2015年6月5日由中国医药科技出版社出版，由国家药典委员会创作
中国药品通用名称	尚未有通用译名的名词术语	是中国法定的药物名称，由国家药典委员会负责制定
经穴部位、耳穴名称与部位	经络针灸学名词术语	其著录方式参照采用世界卫生组织总部针灸穴名国际标准化科学组会议审定通过的《标准针灸穴名》

3 中医诊疗数据集构建方法

3.1 数据集构建流程

出于对数据集应用的综合考虑，在构建过程中应注意以下几点问题：①此数据的主要功能是记录详细的诊疗过程，因此元数据数量较多，要求对数据表的字段划分合理、内容简洁明了、逻辑层次清晰；②为了更好地形成数据共享生态系统，应对数据集进行结构化标记，以达到数据共建共享、知识交流的目的；③由于

此数据集的一个应用场景为实地采集数据，因此数据集采集界面的设计以及后续数据管理系统的开发也是数据集完善和维护需要考虑的重要问题。结合前期准备工作，整体数据集构建流程如图 2 所示：

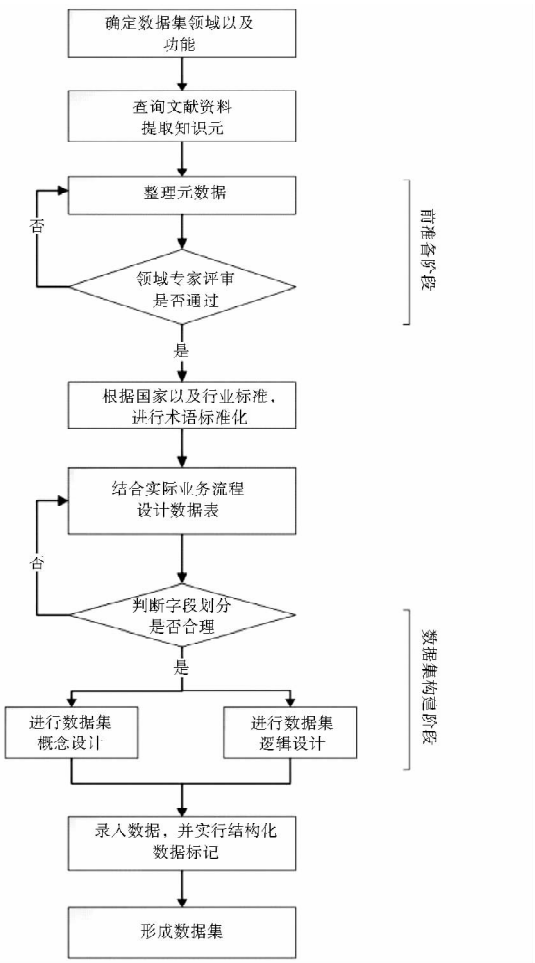

图 2 数据集构建流程

3.2 数据集结构设计

数据集的结构类似于关系数据库的结构，是由表、行、列等对象构成的层次结构，同时也包含数据集定义的约束和关系。因此此处参考关系数据库的构建方法和步骤设计数据集结构。

3.2.1 数据集的概念结构设计

数据库构建有三范式，为数据库的设计提供了基础的逻辑要求，在此基础上根据具体情况作出适当修改。

第一范式为“每个实体的元组中的每一个属性都不可再分”。在上述设计数据表中大部分的列都满足不可分割属性，但在患者基本信息子集中的“辅助检查”字段考虑到患者进行辅助检查的方法众多，每种方法又有各自的专业术语进行描述，是一个十分庞大且复杂的字段，本着录入数据的简便性以及查找数据的便捷性，此处不做穷尽列举，将所有辅助检查信息合并到同一列；同样地，在诊断治法信息子集的“治疗方法”字段中，每种用药的名称、剂量、使用方法以及注意事项均为一一对应关系，不应该把名称或剂量单独提取在一列中，这样就失去了数据的价值，因而在此处，对“方药名称”等元数据进行合并。

第二范式为“每个实体的元组中不能存在与主键字无关的属性”。考虑到对个人隐私的保护，医案中“姓名”等涉及个人隐私的字段常常缺省，因此，此类字段不适合设置为数据表的主键，故在每张数据表中添加“医案编号”字段并设为主键，每张数据表处于并列等级。

第三范式为“每个实体元组中非关键字属性不存在依赖关系”。每张数据的每一列都具有独立性，且划分符合医案基本逻辑，避免了数据冗余和信息混乱。

根据上述规则创建各数据表，并根据实际情况以及数据要求对表中的数据进行约束：

（1）医案标识信息表。如表 3 所示：

表 3 医案标识信息表

列名	数据类型	是否允许空值	备注	标准
医案编号	int	不允许	医案在知识库中的编号	DC
医案名称	nvarchar(20)	允许	医案的标题	DC
医案来源	nvarchar(20)	允许	医案出自哪本医书或病例集	DC
医案作者	nchar(10)	允许	医案创作者或就诊医师姓名	DC
就诊日期	date	允许	患者就诊日期	DC
收录日期	date	不允许	医案收录日期	DC
医案科别	nchar(10)	允许	医案所属中医类别	GB/T 15657 – 1995
病证分类	nvarchar(20)	允许	医案记录患者主病的症候	标准

（2）患者基本信息表。患者基本信息表详细记录患者的个人信息，包括患者姓名、性别等基本信息，也包括既往病史、个人生活史等对病情可能产生影响的因素，其数据标准参考 WS445. 11 – 2014 中华人民共和国卫生行业标准 – 中医病案住院首页部分，此标准详细列出了 GB/T 2261 – 2003 个人基本信息分类与代码、卫生信息数据元值域与代码等，一共分为 17 个部分，此处主要参考第 11 部分。见表 4。

（3）诊疗过程信息表。诊疗过程信息子集记录临床疾病的诊断过程，传统中医通过“望”“闻”“问”“切”四诊合参来了解病情，期间辅以患者主诉情况，医者综合刻下症（即患者就诊时的症状）与患者患病过程来判定病情。此类对病情的主观描述具有很强的

表 4　患者基本信息表

列名	数据类型	是否允许空值	备注	标准
医案编号	int	不允许	外键,引用医案基本信息表	
姓名	nchar(10)	允许	多为患者挂号时登录的个人信息,为保护个人隐私,数据集中可缺省	
体温	nchar(10)	允许	诊室内即可检查的项目,且就诊时多需要量取当下数据	GB/T 2261－2003
辅助检查	nvarchar(50)	允许	各类项目检查结果,例如:体检单,CT、核磁	
过敏史	nvarchar(20)	允许	记录患者既往疾病,重点是慢性病、传染病等	WS445.11－2014
生活习惯	nvarchar(20)	允许	记录与病症相关的个人经历	WS445.11－2014

灵活性,尤其患者主诉部分需要医者通过临床经验转化为医学相关信息,在记录中主要参考两项标准:GB_T 15657－1995 中医病证分类与代码、GB_T 16751.2－1997 中医临床诊疗术语症候部分,此处列举部分数据的示例以便参考。在对主要病因的解读与患者主证的记录方面,不同医案记录格式有所不同,贾李蓉等[21]在关于中医病证分类体系的研究中,根据临床实用性和证候概念的自身特点,形成了证候类概念多维度归类的原则。本次数据集设计具体数据的记录方式可在参考 GB/T 15657－1995 中医病证分类与代码、医疗机构诊疗科目名录－中医科、GB/T 16751.2－1997 中医临床诊疗术语症候部分的基础上进行灵活扩展。如表 5 所示:

表 5　诊疗过程信息表

列名	数据类型	是否允许空值	说明	示例
医案编号	int	不允许	外键,引用医案基本信息表	
主要症状	text	不允许	记录患者主要症状、患病部位及程度	恶心呕吐
舌诊	nchar(20)	允许	主要记录舌色、舌苔	苔白、色红
面色	nvarchar(50)	允许	主要记录面部状态、色泽	面色憔悴、水肿
问寒热	nvarchar(50)	允许	记录患者畏寒或畏热的情况	畏寒、畏风
问汗	nvarchar(50)	允许	记录患者日常以及夜间出汗情况	夜间盗汗
闻声息	nvarchar(50)	允许	记录患者说话、喘息声音	气喘
闻气味	nvarchar(50)	允许	记录患者发出的不正常气味	口臭
脉诊	nchar(20)	允许	中医把脉情况	脉滑、脉细数
按诊	nvarchar(50)	允许	按压患者体表穴位的情况	腹软

(4)诊断治法信息表。诊断治法信息子集主要记录患者主证、使用方剂名称和用法以及一些医者给出的生活习惯建议,参考标准为 GB/T 16751.3－1997 中医临床诊疗术语治法部分。如表 6 所示:

表 6　诊断治法信息表

列名	数据类型	是否允许空值	说明	示例
医案编号	int	不允许	外键,引用医案基本信息表	
患者主证	nvarchar(50)	不允许	患者的主要病症,包含患病部位、患病程度等信息	咳嗽
辩证分析	nvarchar(50)	允许	辨清疾病的病因、性质、部位	风寒证
治则	nvarchar(50)	允许	中医治疗疾病的法则,包含治疗原则与方法	清利湿热、疏肝理气、活血镇痛
方药名称、剂量、用法	nvarchar(100)	允许		
其他医嘱	nvarchar(50)	允许		

(5)影像检查信息表。在患者问诊过程中,可能会提供一些以往影像检查的资料,在医生诊断时,也可能需要相关影像检查信息作为参考,影像检查信息子集详细记录影像检查情况。如表 7 所示:

表 7　影像检查信息表

列名	数据类型	是否允许空值	说明	举例
医案编号	int	不允许	外键,引用医案基本信息表	
影像检查号	nchar(10)	允许	影像科室标识受检者接受某次检查时的特征编号	
检查唯一标识符	nchar(10)	允许	某次检查中影像设备生成的检查唯一标识符	
检查部位	nchar(10)	允许	受检者在某次检查中的检查部位	头部
检查方法	nvarchar(50)	允许	对受检者采用的检查方法	CT
检查日期	nchar(10)	允许		
检查机构	nchar(10)	允许	检查的医疗机构名称	

(6)医案流转信息表。医案流转信息子集记录医案的录入与导出,明确医案的责任人与流动路径。如表 8 所示:

表 8 医案流转信息表

列名	数据类型	是否允许空值	说明
医案编号	int	不允许	外键,引用医案基本信息表
录入人员	nchar(10)	允许	负责将医案录入数据库的人员
录入日期	date	允许	医案入库日期
传递人员	nchar(10)	允许	
传递日期	date	允许	
传递机构	nchar(10)	允许	
责任人	nchar(10)	允许	

(7)医案管理信息子集。医案管理信息子集记录医案的客观管理情况,是对医案内容的客观评价。如表 9 所示:

表 9 医案管理信息表

列名	数据类型	是否允许空值 默认值	说明
医案编号	nchar(10)	不允许	外键,引用医案基本信息表
整理人员	nchar(10)	允许	负责对数据库内医案进行整理的人员
结构化整理	nchar(10)	允许	记录语句、概念粒度等
审核	nchar(10)	允许	
更新	nchar(10)	允许	

以上是对中医诊疗数据集各个信息表的整理,在满足实际实用的基础上尽可能符合国家以及行业标准。中医在不断的发展中会有新概念的产生与旧概念的剔除,故以上所采用的标准也会随着实际应用情况的变化进行更新。

3.2.2 数据集的逻辑结构设计

数据库结构是指在计算机的存储设备上存放合理的、相互关联、有逻辑结构的数据集合的结构。一个数据库结构有很多层次,如数据表、字段等。由于中医医案信息层次众多、相互依存的特点,在对其进行统计分析时,既要从整体角度考虑又要把握每部分的重点信息。如图 3 所示,本次设计数据库结构的特点,就是以“医案编号”为主键,联系起从诊疗过程中医患两方的详细信息再到诊疗结束后保存管理机制的信息。

如此设计避免了一张数据表内信息超载、字段冗长且不易定位的缺点,能让数据分析人员清晰地了解数据集结构,方便进行数据存取与处理。

3.3 数据集采集系统设计

本研究采用开发 c#应用程序来进行中医诊疗数据采集系统的设计和应用,创建 windows 窗体的组件,并利用 DataAdapter 对象实现数据库内容的显示与交互式更新。DataAdapter 对象在数据源与 DataSet 之间起到桥梁作用,既可以将数据录入到 DataSet 的数据表中,也可以将 DataSet 的数据更改送回数据源,这就使采集系统除了可以进行数据采集之外,还包含存储、修改等功能,而数据分析则可在数据库内进行。

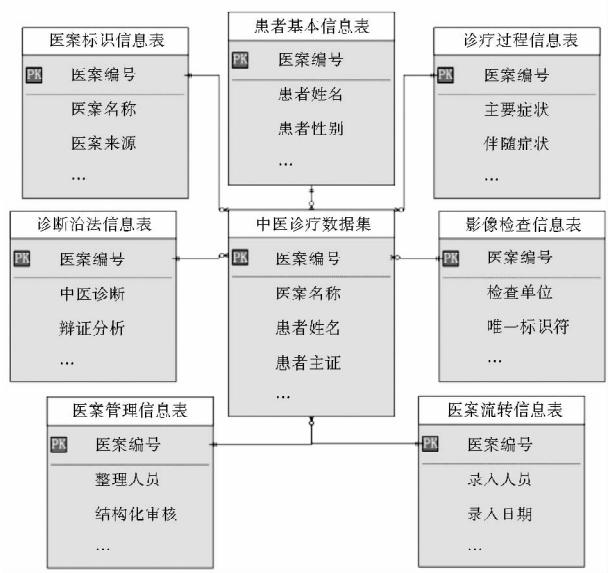

图 3 数据集逻辑结构

对于中医实地诊疗数据采集来说,数据集录入界面设计及功能设计非常重要,对易用性的要求很高,须构建一个友好的信息录入页面,字段安排应符合逻辑,录入时不用反复切换数据表,录入框应设置常用自动选择项,节约录入时间,为医生、患者以及数据导入人员提供便利。在保证录入效率的同时可以将一些非医学字段交由患者自行填写,设计多端数据同步,提高问诊效率,将四诊的具体字段按顺序列出,医生可以在相应位置直接录入。在录入药方以及医嘱时,可以先将药品名称以及用药要求提前录入数据库,医生可以根据实际情况选择默认或更改用药要求,不用再反复输入。数据采集系统的界面见图 4。

4 实证研究

数据集的优点在于能实现高效率的数据访问和操作,且具有快速定位、检索查询、数据交互等功能,为了考量数据集构建的合理性与灵活性,现从网络已有中医医案与实地诊疗两方面采集数据,初步形成数据集,进行实证研究。

4.1 网络医案数据的采集

4.1.1 网络医案数据的筛选

在网络上查找中医医案的过程中发现,现阶段,网

图 4　数据采集系统界面

络上的医学网站数量众多，有涉及中医诊疗方面的数据大多分散，不好爬取，在对中医医案、病例的记载中，不同网页的数据风格迥异，所采用的字段各不相同。考虑到录入数据质量，对网络医案数据筛选提出如下几点原则：

（1）缺乏基本患者信息的不收录。对于患者性别、年龄等必要信息缺失的不予收录。患者的年龄、性别对中医诊断具有重要的参考作用，因此这类信息不可缺失。

考虑到患者的隐私保护问题，网络上中医病例的记载大多以“患者，男，33 岁”或“患者，中年女性”描述，中医崇尚以人为本，同类症状对不同患者而言具有不同的诊疗方法，如果缺乏此类信息，那么后续的诊疗用药描述则不具有分析意义。

（2）缺乏主要症状描述的不收录。在症状表述部分，因为没有统一标准，没有固定格式，所以每个网站、每种疾病都有不同的表述方法，甚至对于同一种疾病下的症状的表述都不尽相同。现阶段网站收集病例为了避免要求过于严苛而无法上传数据的现象，把现病史部分统称为“主诉”或“主要症状”，在此次数据试录的过程中，也主要将这一部分的数据收集进“患者基本信息子集”和“诊疗过程信息子集”。在数据筛选的过程中，剔除掉症状表述模糊不清、过于笼统、存在疑问的病例，以免影响后续数据处理。

（3）缺乏治则的不收录。每一个完整的医案与经方的重点便是治法治则，此类信息是对病例的总结，也是最具统计研究意义的部分，有些网站为了表述中医症状，收集了大量医书和古籍对于该病证的介绍，用于普及和解释一些难以理解的中医病症术语，但其最终治法却缺省或表述不清，对于此类中医症状数据不予收录。

（4）剔除非医案数据。现阶段，大量医学咨询网站上内容丰富，信息繁杂，给网络问诊的患者尽可能地提供更多的帮助，因此在相关病例的记载中，也会记录大量与中医病案无关的字段，如“此病是否属于医保”“最佳就诊时间”“就诊前准备”“三甲医院治疗费用”等诸多信息，此类信息对于患者问诊有很大帮助但与医案记录无关，故应剔除。

（5）存在知识产权限制的不收录。有的网站信息是公开共享的，有的网站标注转载需要告知，而有的是明确要求不可以转发拷贝的，在录入网络数据时应提前了解此方面信息，尊重个人隐私，尊重知识产权，存在知识产权限制的医案不予收录。

4.1.2　数据的试录与评价

网络数据试录的主要目的是检验数据集的合理

性,要求录入数据符合规范,本次数据试录的数据选自“爱爱医医学网——精选中医病例”以及“中医世家——医案心得”两个开源医学网站,这两个网站中医案信息的录入较为全面,且字段划分与本研究相近,前者采用手工整理的方式采集 54 条精选中医医案数据,后者在剔除不相关信息后,手工整理数据 47 条,总共整理数据 101 条,均保存为 Excel 格式。

在试录数据的过程中可以发现,网络上的医案数据虽然总量繁多,但存在一些数据存疑(经网站专家评论指出)、信息表述模糊、数据缺省过多等现象,数据利用率较低。且每一个网站上记载医案信息的字段都不尽相同,且收录侧重点不一,基本上是一个“字段”对应一个或多个“信息表”的情况,给数据清洗带来极大困难,很难进行批量处理。但此次试录也证实了网络医案中的数据在此次构建的数据集中均有其匹配字段,不存在已有医案相关数据却无字段可存放的现象,这也证明本研究的数据集设计可以满足收录网络中大部分医案病例的需求。

4.2 中医实地诊疗数据采集

录入实地诊疗过程所产生的数据相比于网络导入数据更能体现此数据集的优点。由于中医问诊特点,在实地诊疗过程中,大量的隐性数据无法收集,比如“望神态”“舌诊”“脉诊”等,中医在诊疗过程中常常从细枝末节的地方收集患者与病情相关的信息,但却由于看病效率等问题不会记录在医案中,因此导致数据的流失。

4.2.1 实地采集方法

在实地采集数据前,应对所需要研究的数据提前做好评估,规划数据数量以及门诊科室等信息,在实地采集中可以采用多种方法以完成目标任务。

(1)由医生本人录入。医生在诊疗过程中,一边问诊,一边在数据采集界面录入数据,在问诊过程中,对诊疗具体病症所涉及到的元数据信息尽可能全面录入。医生对诊疗过程了解最为深入,这种方法的优点是可以最大程度地保证数据的质量,缺点是会影响医生问诊效率,对于患者数量较多的医生来说存在一定的限制。

(2)由医生助理录入。具有一定资历的中医一般都有医生助理辅助进行诊疗管理,可以由医生助理完成录入工作。诊疗过程中,医生需要随时将望闻问切得到的信息口述出来,由医生助理完成信息录入,数据的审核整理也可以同步进行,更高效率地利用诊疗时间。

(3)研究人员录入。在数据集创建过程中,可以由项目研究人员完成录入。一种方法是进行现场录音,完全不干扰医生问诊,后期根据录音完成数据录入;另一种方法是研究人员在现场根据医生问诊的情况及时录入。研究人员录入的优点是不影响医生问诊效率,而且研究人员对于录入系统的使用比较熟练;缺点是如果采用录音的方法,需要在录音前做好患者隐私保护方面的沟通;在录入完成后,需要由医生或者医生助理进行医案审核。无论是录音还是现场录入,都需要诊疗医生将望闻问切等元数据涉及到的信息口述出来。

(4)多种方法融合采集。因为实地采集具有很高的灵活性,所以在问诊现场可以同时采用多种方法,安排不同人员分工合作,从不同角度对实时产生的数据进行采集,达到最高的数据准确度,减少数据遗漏。但采集过程需要提前模拟演练,在采集时注意配合,不影响到问诊效率。

4.2.2 实地数据采集

中医诊疗数据的采集非一日之功,既不能影响医生正常问诊效率,又要保障数据集数据量,因此需要日积月累、逐步完善和充实数据集。

进行实地数据采集之前,先由本次研究人员进行模拟问诊,提高录入数据的熟练度。实地数据采集时,采用多种方法融合采集,由研究人员与医生及医生助理共同配合,研究人员进行患者基本信息以及最终药方医嘱的录入;医生则根据此次数据集设计,在诊疗过程中将元数据涉及到的信息尽可能地口述出来;医生助理录入四诊数据,校准专业医学术语。并在不泄露患者个人隐私的情况下进行现场录音,以保证能够核实数据的全面性和准确性。相较于网络数据爬取,现场采集的数据在数据质量上得到了保证。

4.3 数据集结构化数据标记

构建数据集并非闭门造车,一个优秀的数据集也不能一蹴而就,而是要不断通过实际检验来修改和完善。现如今不同领域建立起的数据集不胜枚举、种类繁多,给检索和查找带来了一定难度,因此也产生了专业的科学数据集检索平台,既有综合性的检索平台,如 DCI、Dataset Search;也有聚焦于生物医学领域的 DataMed 等。为了更好地检验研究成果,越来越多的数据集存储区采用 schema. org 及类似标准来描述数据集,用户通过数据集搜索找到的数据集的种类和覆盖范围将会持续增加,这样更利于数据共享共建。

考虑到数据集后续应用,此处采用 Google 的数据

集发现方法，将 schema. org 标准添加到描述数据集的网页，并通过结构化测试工具来验证数据集网页已添加的结构化标记。

上传后的数据集如图 5 所示，在后期数据采集过程中，除本团队研究人员之外，也会邀请协作者共同完善数据集，进一步扩大数据量。

图 5　数据集展示界面

4.4　数据应用展望

秘红英等[22]在第十四届国际络病学大会上归纳了几类中医医案的分析方法，除了个人领悟分析之外，还可以运用统计学的方法对医案集内数据进行处理，相比于个人领悟，统计的方法有其系统性、科学性的优势。张晓航等[23]探讨了机器学习方法与深度学习方法在中医诊疗中的应用，传统机器学习算法如聚类算法、分类算法、回归分析算法、关联规则算法在中医领域已有较多应用，深度学习算法则更接近中医内核，是中医未来发展的大方向，而机器学习算法的成功，离不开大量高质量数据的支持。此次数据集构建将重点放在每位患者自身的情况记录与分析上，弥补了现有数据集只专注于某一疾病或忽视患者体质的问题，为中医领域的数据分析提供了重要支撑。

5　结语

科学的数据分析是医疗卫生事业发展的主要方向，然而在中医领域却存在数据流失与精确度不够等问题。本次研究通过构建基于元数据的中医诊疗数据集的方式，详细记录中医四诊合参、辩证分析的过程，形成一个结构化的数据体系，打破信息孤岛，将患者、医师以及研究人员都纳入到数据记录整理的环节中来，促进中医诊疗知识的保存、传播与再利用，再通过数据库应用技术对数据进行分析与更深层次的挖掘，达到知识共享与传承的目的。但本次研究仍存在不足之处：①应用自然语言处理技术来增强对医案古籍以及现代名家医案的批量处理能力；②面对复杂病例的记载，应更好地体现时间维度，在详细记录患者多年病情的同时尽量减少信息冗余。在未来的研究中，将以本文所构建数据集为基础，在不断的实证中积累经验、优化采集方案、完善数据管理系统，为中医诊疗知识组织工作提供更全面的辅助支撑。

参考文献：

[1] 王丹丹. 科学数据出版过程中的数据质量控制[J]. 图书情报工作,2015,59(23):124－129.
[2] 鲁兆麟. 中华历代名医医案全库[M]. 北京:北京科学技术出版社,2015.
[3] 王佑华,陆金根,柳涛,等. 中医医案中的知识发现研究[J]. 中西医结合学报,2007,5(4):368－372.
[4] 刘丹红,张林,杨喆,等. 医学语言与临床数据标准化概述[J]. 中国卫生信息管理杂志,2014,11(1):14－17.
[5] 詹林城,苏华冠,杨伟锋,等. 病案首页数据质控知识库建立与应用探讨[J]. 医学信息学杂志,2019,40(5):72－76.
[6] 徐坤. 基于本体的科学数据监护平台研究[D]. 长春:吉林大学,2014.
[7] 贾李蓉,聂莹,李海燕. DC、CKRM、TCMLM 的比较研究[J]. 中华医学图书情报杂志,2018,27(12):36－42.
[8] 赵阳,崔蒙. 中医文献元数据设计原则和实用性思考[J]. 世界科学技术－中医药现代化,2015,17(10):1978－1981.
[9] 刘宝杰,陈进,马路. 中医病案数据库元数据结构设计实践[J]. 医学信息学杂志,2013,34(12):70－73.
[10] 孙静,杨帆,邓文萍,等. 基于本体的中医症状知识表示模型构建[J]. 医学信息学杂志,2017,38(2):52－56.

[11] 刘丽华,胡凯,金水高. 卫生信息数据集元数据规范的研究[J]. 中国卫生统计,2008(4):363-365,372.
[12] 刘敏娟,张学福,颜蕴,等. 基于期刊主题相似性的领域分析数据集构建:方法与实证[J]. 图书情报工作,2016,60(10):115-122.
[13] 单连慧,李勇,李海存,等. 基于DOI构建医学领域科技评价文献数据集的方法研究[J]. 医学信息学杂志,2013,34(2):35-39,44.
[14] 邱瑞瑾,李敏,胡嘉元,等. 中成药上市后临床安全性评价核心数据集的构建方法探索[J]. 世界科学技术-中医药现代化,2018,20(10):1723-1728.
[15] 李姣,郭海红,郭珉江,等. 美英政府开放健康医疗数据的主题分布与开放程度量化研究[J]. 图书情报工作,2015,59(20):132-137.
[16] DING E,ALBUQUERQUE D,WINTER M,et al. Abstract 15983: novel method of atrial fibrillation case identification and burden estimation using data in the electronic health record: data from the MIMIC-III dataset[J]. Circulation,2018,138(S1):A15983.
[17] 曾蕾. 元数据与专业置标语言在数字图书馆中知识表述方面的功能[J]. 图书情报工作,2002,46(10):14-22.
[18] CAMPBELL A J,FABRIZIO B,W L H,et al. Speaking the same language: using standardized terminology[J]. Journal of lower genital tract disease,2016,20(1):8-10.
[19] 李璐. 面向中医诊疗知识库的医案元数据模型构建研究[D]. 长春:吉林大学,2020.
[20] 冯磊. ISO首发两项中医药信息国际标准[N]. 中国中医药报,2014-08-27(1).
[21] 贾李蓉,刘静,刘丽红,等. 中医临床术语系统v2.0病证分类体系构建研究[J]. 中国中医药图书情报杂志,2018,42(5):8-12.
[22] 秘红英. 中医医案的分析方法[C]//中国工程院医药卫生学部,中华中医药学会,世界中医药学会联合会,等. 第十四届国际络病学大会论文集. 济南:中华中医药学会络病分会,2018:240-244.
[23] 张晓航,石清磊,王斌,等. 机器学习算法在中医诊疗中的研究综述[J]. 计算机科学,2018,45(S2):32-36.

作者贡献说明:
马捷:论文选题制定,研究框架设计,论文修改及定稿;
王珏:数据集结构设计,论文撰写及修改;
孙恒宇:中医领域专业知识的审核与修改;
李璐:元数据的调研及拟定。

Data Set Construction of TCM Diagnosis and Treatment Based on Medical Case Metadata: Methods and Empirical Evidence

Ma Jie[1,2] Wang Jue[1] Sun Hengyu[3] Li Lu[1]
[1] School of Management, Jilin University, Changchun 130022
[2] Information Resources Research Center of Jilin University, Changchun 130022
[3] The First Clinical Hospital of Jilin Academy of Traditional Chinese Medicine, Changchun 130022

Abstract: [**Purpose/significance**] TCM medical case is an important carrier to record the diagnosis and treatment data. It is of great significance for TCM diagnosis and treatment knowledge sharing, analysis and inheritance to collect medical case data based on metadata and form a data set. [**Method/process**] We selected and determined of TCM medical case metadata, and referred to the relational database for the conceptual and logical design of the data set, collected, organized and stored TCM medical case data based on case description to form the method of data set. [**Result/conclusion**] Network and field diagnosis and treatment data collection were conducted based on relatively complete TCM medical case metadata. The formed TCM diagnosis and treatment data set can not only support multi-dimensional diagnosis and treatment information retrieval, but also provide data sources for further diagnosis and treatment data analysis and mining.

Keywords: medical record TCM medical record metadata TCM diagnosis and treatment dataset